DNMAD: web-based diagnosis and normalization for microarray data

Juan M. Vaquerizas, Joaquín Dopazo and Ramón Díaz-Uriarte*

Bioinformatics Unit, CNIO, Spanish National Cancer Centre, Melchor Fernández Almagro 3, 28029 Madrid, Spain

Received and revised on May 31, 2004; accepted on July 1, 2004
Advance Access publication July 9, 2004

ABSTRACT
Summary: We present a web server for Diagnosis and Normalization of MicroArray Data (DNMAD). DNMAD includes several common data transformations such as spatial and global robust local regression or multiple slide normalization, and allows for detecting several kinds of errors that result from the manipulation and the image analysis of the arrays. This tool offers a user-friendly interface, and is completely integrated within the Gene Expression Pattern Analysis Suite (GEPAS).

Availability: The tool is accessible on-line at http://dnmad.bioinfo.cnio.es
Contact: rdiaz@cnio.es

1 INTRODUCTION
DNA array technologies (Schena et al., 1996) allow the simultaneous monitoring of the expression of thousands of genes. Experiments usually involve measuring the expression levels of genes under several conditions corresponding to different samples, different doses of a certain drug or different stages during an individual’s lifetime.

Ideally, DNA microarray techniques allow us to focus on changes in gene expression levels that are solely due to differences between the samples. However, it is widely accepted that other factors such as differences in labelling efficiency, physical properties of the dyes, differences in hybridization, changes in scanner settings, etc. can introduce systematic variation aside from that due to the expression levels of genes in each sample (Yang et al., 2002).

Therefore, to obtain accurate and precise results from the analysis of microarray data, normalization of data is essential. Here, we describe a web-based tool for Diagnosis and Normalization of spotted cDNA MicroArray Data (DNMAD).

2 MOTIVATION
Within print-tip group location normalization has proven to be an effective method in normalizing microarray data (Yang et al., 2002), but the use of most currently available implementations requires prior training on using a statistical package such as R (http://www.R-project.org), which could represent a relatively high investment especially if the only aim behind such training is to normalize microarray data.

Despite the importance of normalization, there are not many bioinformatic Web tools available to perform normalization procedures, and the few that exist, e.g. SNOMAD (Colantuoni et al., 2002) or Multi microarray normalization (http://genome1.beatson.gla.ac.uk/Rweb/anova.html), do not offer certain options that would make the process more efficient such as the use of files coming directly from the scanner or the possibility of entering more than one slide at a time.

In addition, connectivity and data exchange between tools is also important. Web tools integrated within a suite for gene expression pattern analysis, as opposed to individual Web tools, allow users to perform the whole analysis without having to manipulate file formats to make them compatible with the tools used for each individual step. Also, server-based solutions benefit from the server processing capacity (greater than that of personal computers), that allow the user to process more data simultaneously. To allow for efficient handling of requests from many users, the DNMAD service is controlled by a queue system.

The aim of this tool is to provide the scientific community with an easy to use, fully featured web interface for microarray data normalization that is completely integrated within the freely available Gene Expression Pattern Analysis Suite (GEPAS) (Herrero et al., 2003a).

3 WEB INTERFACE
This application uses R (R Core, 2004, http://www.R-project.org) and the Bioconductor package limma (Smyth et al., 2004, http://www.bioconductor.org), with some custom modifications, to carry out the normalization procedures. The web interface is a Perl CGI script that communicates with R using the CGIwithR (Firth, 2004, http://cran.r-project.org) package. The server accepts a set of GenePix files as input, either compressed as a single file (.tar.gz, .zip or .tar.bz2) or as uncompressed GenePix files coming directly from the scanner. These files must adhere to the original standard file format from GenePix, although customized files containing only the
appropriate columns of data (Block, Column, Row, Name, ID, F635 Mean, B635 Median, F532 Mean, B532 Median and Flags) can also be used if GenePix software has not been used to scan the arrays, or if custom modifications have been made to the data. These columns constitute the minimal set of columns required to perform the normalization. The layout of the array must be introduced into the interface along with the data files. Other options, such as the use of flags (for excluding certain spots from the normalization) or background subtraction, can also be selected in the web interface.

The default normalization method is ‘print-tip loess’, as implemented in Yang et al. (2002). This method consists of a robust local regression for each print-tip group of the array. This method can only be used if the following assumptions are met: (1) the number of points to normalize must be large in each print-tip group; (2) very few genes should be differentially expressed (i.e. the expression of the two co-hybridized mRNA samples should be similar for the majority of the spots of the array); and (3) there should be an approximately equal number of up- and downregulated genes in each print-tip group.

If these assumptions cannot be met for each print-tip group, it is possible to use, instead, ‘global loess’, where the local regression is carried out over the whole array instead of for each print-tip group (Yang et al., 2002).

4 RESULTS

The output page contains information regarding the type of normalization followed, errors and warnings, plots for interpreting the results and for diagnosis and links to download the normalized data or to enter the hub of GEPAS (Herrero et al., 2003a), via the Preprocessor module (Herrero et al., 2003b).

A summary of all options selected for normalization is shown at the top of the output page. This information is also attached to the results file that contains the normalized log ratios of expression. Furthermore, a collection of plots and images are provided in order to show plate effect, printing effects and potential problems occurring during the elaboration of the array and the hybridizations, in the sample preparations, in the scanning of the arrays, etc.

These plots include: (1) boxplots for all the arrays and for each array showing individual print-tip groups, to assess the need for normalization for a particular array; (2) MA-plots with the regression curve for each print-tip group, in order to observe the efficiency of the normalization; and (3) diagnostic plots.

Diagnostic plots are compositions of 10 different plots, and they should help the user to detect problems in the arrays (Smyth et al., 2003, http://www.bioconductor.org). These plots include: (1) histograms of the raw pixel intensities (log$_2$) of the red and green mean foreground that help to identify problems in the scanner settings or in the hybridizations; (2) density plots of the log$_2$ intensities for both channels, displayed in sidewise panels for all the arrays and for each array; and (3) images of the arrays including the red and green background, and the unnormalized and normalized ratio values, which should help to identify spot damaged arrays or spatial patterns. All plots and images can be downloaded along with an HTML document to browse through them.

As the server accepts multiple arrays, and given that these multiples arrays can show differences between their scales, slide-scale normalization is provided. This method, implemented as in Yang et al. (2002), reduces differences in the scales of the arrays that are being normalized. After slide-scale normalization, plots and images are provided again, allowing the user to decide if the slide-scale procedure is valuable for the particular data.

Finally, links to download the normalized data [normalized expression ratios in log$_2$ scale and the A values (the ‘average signal’ or 0.5 + (log$_2$ R + log$_2$ G)) are displayed, to enable data storage. A direct link to the Pre-analysis module of the Preprocessor (Herrero et al., 2003b) is also provided, allowing the user to continue with analysis using the GEPAS suite (Herrero et al., 2003a, 2004), such as for clustering genes, or to identify differentially expressed genes, without having to perform any kind of format adjustment to the data.

ACKNOWLEDGEMENTS

We thank A.Lee and an anonymous reviewer for comments on the manuscript, A.Wren for revising the English of the manuscript. We are grateful to S.Dudoit, G.Smyth and P.Kemmeren who provided answers and suggestions about normalization and diagnostics. This tool would not have been possible without the excellent, and free, statistical computing system R and several of its packages available from CRAN or part of Bioconductor, all developed and maintained by volunteers. R.D.-U. was partially supported by a Ramón y Cajal programme from the Spanish Ministry of Science (MCyT). This work is also partly supported by Fundación BBVA and by Project TIC2003-09331-C02-02 of the Spanish MCyT.

REFERENCES


